

Entropy based feature selection for text categorization

Christine Largeron
Université de Lyon, F-42023,
Saint-Étienne, France
CNRS UMR 5516, Laboratoire
Hubert Curien
christine.largeron@univ-
st-etienne.fr

Christophe Moulin
Université de Lyon, F-42023,
Saint-Étienne, France
CNRS UMR 5516, Laboratoire
Hubert Curien
christophe.moulin@univ-
st-etienne.fr

Mathias Géry
Université de Lyon, F-42023,
Saint-Étienne, France
CNRS UMR 5516, Laboratoire
Hubert Curien
mathias.gery@univ-st-
etienne.fr

ABSTRACT

In text categorization, feature selection can be essential not only for reducing the index size but also for improving the performance of the classifier. In this article¹, we propose a feature selection criterion, called *Entropy based Category Coverage Difference (ECCD)*. On the one hand, this criterion is based on the distribution of the documents containing the term in the categories, but on the other hand, it takes into account its entropy. *ECCD* compares favorably with usual feature selection methods based on document frequency (*DF*), information gain (*IG*), mutual information (*IM*), χ^2 , *odd ratio* and *GSS* on a large collection of XML documents from Wikipedia encyclopedia. Moreover, this comparative study confirms the effectiveness of selection feature techniques derived from the χ^2 statistics.

1. INTRODUCTION

Text categorization (or classification) is a supervised task for which, given a set of categories, a training set of preclassified documents is provided. Given this training set, the task consists in learning the class descriptions in order to be able to classify a new document in one of the categories ([25, 21, 19]).

Irrespective of the categorization method applied, the documents must be represented by a set of features. Several models can be used, such as the boolean model, the probabilistic model or the vector space model which described documents as a bag of words. As each term belonging to at least one document can be considered as a feature, this can lead to an index size (*i.e.* the feature space dimension) very large, even for a small collection composed of short documents like news articles. Moreover, all these words are not equally useful for the categorization, like for instance stop words, synonymous, *etc.* Their distribution must also be

studied. For example, words that appear in a single document or in all the documents are not relevant for the categorization task. So, we need to extract a more effective set of features from the text, that can be used to efficiently represent the documents for their categorization. This preprocessing step, can be essential on textual data for improving the performance of the categorization algorithm ([26, 7]).

Among the many methods that can be used, two strategies can be distinguished: on the one hand, the dimension reduction which consists in creating new synthetic features that are a combination of the original ones, like for instance the Latent Semantic Analysis (*LSA*) which uses a Singular Value Decomposition (*SVD*) to extract concepts from the words [2] and, on the other hand, the feature subset selection that is based on non useful features elimination [9]. In this last case, the selection can be embedded, in other words integrated in the categorization algorithm itself. The feature selection can also be based on a wrapper approach which tries different subsets of features as input and takes the subset that produces the best results. Finally, the selection can be done before the categorization, using a criterion to filter the features.

The common feature selection criteria proposed in the literature ([24, 1, 17, 8]) notably, document frequency (*DF*), information gain (*IG*), χ^2 or mutual information (*IM*) consider the distribution of the documents containing the term between the categories but they do not take into account the frequency of the term between the categories. However, we can note that a term which is characteristic of a category must appear in a greater number of documents belonging to this category than into the other categories but it should also appear more frequently. For this reason, we propose the *Entropy based Category Coverage Difference*, denoted (*ECCD*), which exploits also the entropy of the term.

In this article, this *ECCD* criterion is also compared to usual feature selection methods mentioned above (*IG*), χ^2 and mutual information (*IM*)) on a large collection of XML documents. Indeed, previous works have already shown that removing up to 90% of terms can improve the classification accuracy measured by precision [26]. However, these experiments have been performed on corpora composed of short documents such as the well known Reuters Collection, composed of short news articles ([12, 13, 24, 16, 26, 5, 8, 1]). The second aim of this work is to verify the performance of these feature selection methods on the large INEX XML Mining collection composed of heterogeneous XML documents extracted from the Wikipedia encyclopedia [3].

ECCD proposed in this article is defined in section 2 while

¹This work has been partly funded by the Web Intelligence project (région Rhône-Alpes: <http://www.web-intelligence-rhone-alpes.org>)

the usual criteria are detailed in section 3. The experiments and the obtained results are presented respectively in sections 4 and 5.

2. FEATURE SELECTION BASED ON ENTROPY

2.1 Textual document representation

In text categorization, the vector space model (VSM) introduced by Salton *et al.* [20] is widely used as well for flat documents as for semi structured documents written in markup languages like HTML or XML. In this model, documents are represented as vectors which contain term weights. Given a collection D of documents, an index $T = \{t_1, t_2, \dots, t_{|T|}\}$, where $|T|$ denotes the cardinal of T , gives the list of terms (or features) encountered in the documents of D . A document d_i of D is represented by a vector $\vec{d}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,|T|})$ where $w_{i,j}$ represents the weight of the term t_j in the document d_i . In order to calculate this weight, the TF.IDF formula can be used [20]:

$$w_{i,j} = \frac{n_{i,j}}{\sum_l n_{i,l}} \times \log \frac{|D|}{|\{d_i : t_j \in d_i\}|}$$

where $n_{i,j}$ is the number of occurrences of t_j in document d_i normalized by the number of occurrences of all terms in document d_i , $|D|$ is the total number of documents in the corpus and $|\{d_i : t_j \in d_i\}|$ is the number of documents in which the term t_j occurs at least one time.

However, even for limited collections, the dimension of the index can be exceedingly large. For example, in INEX collection, 652,876 non trivial words have been identified. In a small collection of 21,578 documents extracted from Reuters news, more than 40,000 non-trivial words could be identified [19].

Moreover, in the context of categorization, the terms belonging to this bag of words are not necessarily discriminant features of the categories. For this reason, non useful words must be removed, in order to extract a subset T' from T more suited for the categorization task. For that purpose, the local approach consists in filtering a specific subset for each category in such a way that the indexes used to represent documents belonging to different categories are not the same, while the global approach, adopted in this work, uses the same subset T' extracted from T to represent all the documents of the collection ([21, 10]).

In this article, we introduced the ECCD criterion in order to select a subset T' from T , providing a more efficient description of the documents. *ECCD* considers not only the number of documents belonging to the category but also the number of documents belonging to the other categories. Moreover, this criterion also takes into account the entropy of the term. Thus, *ECCD* exploits two hypotheses to evaluate if a term is a characteristic feature for a category. According to the first one, the major part of the documents containing this term must belong to this category. According to the second one, its number of occurrences must be important in the documents of the category and, on the contrary, it must be lower in the other categories.

2.2 Entropy based Category Coverage Difference criterion (ECCD)

Let n_j^k be the number of occurrences of t_j in the category c_k and, tf_j^k the frequency of t_j in this category c_k :

$$tf_j^k = \frac{n_j^k}{\sum_k n_j^k}$$

The Shannon entropy $E(t_j)$ of the term t_j is given by [22]:

$$E(t_j) = - \sum_{k=1}^r (tf_j^k) \times (\log_2(tf_j^k))$$

The entropy is minimal, equals 0, if the term t_j appears only in one category. We consider that this term might have a good discriminatory power in the categorization task. Conversely, the entropy is maximal, equals E_{max} , if t_j is not a good feature to represent the documents *i.e.* if t_j appears in all the categories with the same frequency. *ECCD*(t_j, c_k) is defined by:

$$ECCD(t_j, c_k) = (P(t_j|c_k) - P(t_j|\bar{c}_k)) \times \frac{E_{max} - E(t_j)}{E_{max}} \quad (1)$$

with $P(t_j|c_k)$ (respectively $P(t_j|\bar{c}_k)$) the probability of observing the word t_j in a document belonging to the category c_k (respectively the other categories):

$$P(t_j|c_k) = \frac{|\{d_i \in c_k : t_j \in d_i\}|}{|c_k|}$$

$$P(t_j|\bar{c}_k) = \frac{|\{d_i \notin c_k : t_j \in d_i\}|}{|D| - |c_k|}$$

where $|c_k|$ represents the number of documents in the category c_k and $(|D| - |c_k|)$ the number of documents in the rest of the collection.

We can note that the probabilities of observing the word t_j in c_k and in \bar{c}_k could also be compared with a ratio but this ratio is undefined when the denominator is null. For this reason, it is better to use a difference as done in the formula 1. It is obvious that $P(t_j|c_k)$ increases with the number of documents in c_k containing t_j . Consequently, the higher the number of documents in c_k containing t_j and the lower the number of documents in the other categories containing t_j , the higher the first part of the formula 1. So, t_j is a characteristic feature of the category c_k if the value of *ECCD*(t_j, c_k) is high, in other words, when the major part of the documents containing this term belongs to this category and simultaneously, its number of occurrences is higher in c_k than in the other categories.

Given a term t_j and a category c_k , *ECCD*(t_j, c_k) can be computed from a contingency table. Let A be the number of documents in the category containing t_j ; B , the number of documents in the other categories containing t_j ; C , the number of documents of c_k which do not contain t_j and D , the number of documents in the other categories which do not contain t_j (with $N = A + B + C + D$):

	c_k	\bar{c}_k
t_j	A	B
\bar{t}_j	C	D

Using this contingency table, equation 1 can be estimated

by:

$$ECCD(t_j, c_k) \approx \frac{AD - BC}{(A + C)(B + D)} \times \frac{E_{max} - E(t_j)}{E_{max}}$$

3. COMPARATIVE STUDY OF FEATURE SELECTION CRITERIA

In the next section, we compare the criterion defined in the previous section with other usual selection feature methods which have proved good performances in text categorization. These methods are the document frequency (*DF*), information gain (*IG*), χ^2 , mutual information (*IM*) and *odd ratio*. In our experiments mutual information (*IM*) and *odd ratio* have been less effective. Consequently, only the definitions of the first ones, for which we will detail the results in the next section, are given.

3.1 Document frequency Thresholding (DF)

This method exploits the hypothesis according to which a term belonging to a few number of documents is not a good feature for the categorization task [24].

So, only the terms, that appear in a number of documents higher to a defined threshold, are selected. This threshold can be determined using a training set.

Given a term t_j , this criterion can be computed globally on the collection ($DFG(t_j)$) or on each category c_k ($DFL(t_j, c_k)$):

$$DFG(t_j) = P(t_j) \approx \frac{A + B}{N}$$

$$DFL(t_j, c_k) = P(t_j | c_k) \approx \frac{A}{A + C}$$

One usual way to apply this method consists in eliminating all the words which appear in less than x documents, x varying between 1 and 3 ([24, 5, 15]). Frequently, this technique is used with another feature selection method.

3.2 Information Gain (IG)

Given a term t_j and a category c_k , the information gain $IG(t_j, c_k)$ is defined by [1]:

$$\begin{aligned} IG(t_j, c_k) &= P(t_j, c_k) \log\left(\frac{P(t_j, c_k)}{P(t_j)P(c_k)}\right) \\ &\quad + P(\bar{t}_j, c_k) \log\left(\frac{P(\bar{t}_j, c_k)}{P(\bar{t}_j)P(c_k)}\right) \\ IG(t_j, c_k) &\approx -\frac{A + C}{N} \log\left(\frac{A + C}{N}\right) \\ &\quad + \frac{A}{N} \log\left(\frac{A}{A + B}\right) + \frac{C}{N} \log\left(\frac{C}{C + D}\right) \end{aligned}$$

Only the words for which the value of the criterion is the most important are considered as characteristics for c_k .

3.3 The χ^2 and its extension GSS

χ^2 is used to measure the independence between a term t_j and a category c_k . It is originally defined by:

$$\chi^2(t_j, c_k) = \frac{N \cdot [P(t_j, c_k) \cdot P(\bar{t}_j, \bar{c}_k) - P(\bar{t}_j, c_k) \cdot P(t_j, \bar{c}_k)]^2}{P(t_j) \cdot P(\bar{t}_j) \cdot P(c_k) \cdot P(\bar{c}_k)}$$

$$\chi^2(t_j, c_k) \approx \frac{N \cdot (AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

This criterion equals 0 when t_j and c_k are independent. On the contrary, t_j is considered as a characteristic feature for c_k if the value of $\chi^2(t_j, c_k)$ is high.

Ng *et al.* [17] have observed that the power of 2 at the numerator must be eliminated because it gives the same importance to the probabilities that indicate a positive correlation between t_j and c_k and to those that indicate a negative correlation. Galavotti *et al.* have also proposed to eliminate on the one hand N , that is constant, and on the other hand, the probabilities at the denominator that emphasize very rare features and very rare categories [8]. By removing these factors, Galavotti has introduced the *GSS* criterion:

$$GSS(t_j, c_k) = P(t_j, c_k) \cdot P(\bar{t}_j, \bar{c}_k) - P(\bar{t}_j, c_k) \cdot P(t_j, \bar{c}_k)$$

$$GSS(t_j, c_k) \approx \frac{AD - BC}{N^2}$$

Both χ^2 and *GSS* criteria will be presented in the experiments even if the latest is less popular.

4. EXPERIMENTS

In order to evaluate and compare previously defined selection feature criteria, we performed several experiments on the INEX XML Mining collection [3][4]. After describing this collection, we detail the experimental protocol and we present the obtained results.

4.1 Collection description

The different presented selection feature criteria were subject of several studies in the past. The main used collection for these studies is the well known Reuters-21578 collection proposed by Lewis [14]. This collection is composed of 21,578 documents published on the Reuters newswire in 1987 and manually labelled by Carnegie Group, Inc. and Reuters, Ltd. This collection is often used, although it suffers from many problems. The main drawback is its size. Indeed, the number of documents that composed the collection is small with only 21,578 documents. On top of that, these documents are short with on average 137 words per document. This collection is also disputed in reason of the famous *blah blah blah* [14] that composes some documents. At last, different subsets were used for training and testing in the different studies, that make the comparison tough between obtained results.

In this article, we aim to verify that feature selection, frequently evaluated on small collections of short documents, is also efficient on a bigger collection composed of heterogeneous XML documents. The collection, used in the experiments, is the INEX XML Mining collection from the competition INEX 2008 [4]². This INEX XML Mining collection is composed of 114,366 documents extracted from the well known Wikipedia encyclopedia [3]. The number of words per document is 423 on average, which is three times more than for the documents of Reuters-21578. Moreover, the index size is greater for the INEX XML Mining collection with 160,000 words compared to 40,000 words for the Reuter-21578 collection.

Documents of the INEX XML Mining collection are categorized into 15 categories that correspond to specific topics, such as *sociology*, *sport*, *fiction*, *europe*, *tourism*, *united states*, *etc..* Each document belongs to exactly one category.

²INEX competition: <http://www.inex.otago.ac.nz>

The training set, given by the competition INEX, is created using 10% of the whole collection.

4.2 Preprocessing, feature selection and categorization

Before performing the categorization task, the index of words that is used to represent documents is preprocessed. In order to perform the indexing, the LEMUR software has been used³. Without preprocessing, the original index size is 652,876. In order to reduce this index, we perform a common stemming step using the Porter algorithm [18] that aims to reduce each word to its root form. Thanks to this stemming, the index is reduced to 560,209 distinct words. A lot of words that are not discriminant for the categories are useless. Thus, we remove numbers and words that contains figures (7277, -2311, 0b254c, etc.). The obtained index T after removing all useless words is composed of 161,609 words over the whole collection and 77,706 words in the training set.

However, as explained in the previous section, the terms of T are not necessarily appropriated for the categorization task inasmuch they are not discriminatory for the categories. Thus, thanks to the feature selection criteria, we create a subset T' from T that is more representative than T .

All criteria have been used in the same way: firstly, a number n of words that will be selected for each category, is set. Secondly, given a category, terms are ranked by decreasing order in function of the selection feature criteria values and, the first n words are selected. Finally, the index T' is composed of the union of the first n selected words for each category. Experiments have been done with values of n varying between 10 to 5,000.

In our experiments, the categorization task is performed using SVM algorithm [6] available in the Liblinear library⁴. SVM has been introduced by Vapnik [23] for solving two class pattern recognition problems using Structural Risk Minimization principal. Many studies showed the efficient of SVM in text categorization [11]. The results provided by this approach on the INEX XML Mining collection are presented in the next section.

5. RESULTS

In our experiments mutual information (IM) and *odd ratio* have been less effective. Consequently, only the results obtained with $ECCD$, DF , IG , χ^2 and GSS are given.

Two different measures are used to evaluate the performance of the categorization. The first one is the classification rate which corresponds to the number of correct labelled documents divided by the total number of documents. The second measure is the reduction rate obtained using the index T' relatively to the original index T (77,706 words) and it is computed by $\frac{|T|-|T'|}{|T|}$.

If we consider the original index T composed of 77,706 words and if we perform the categorization, we obtain a classification rate of 78.79%. Figure 1 represents the obtained classification rate in function of the size of the index T' that depends on the number n of words selected by category with the different selection feature criteria. As shown by this figure, reducing significantly the index size leads to

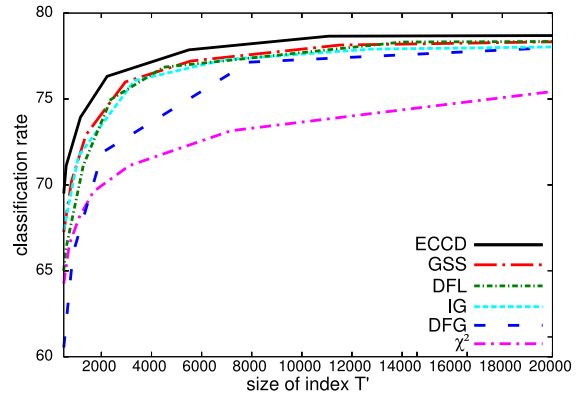


Figure 1: Classification rate obtained for different index size using different feature selection criteria.

still quite good results. Except for the DFG and the χ^2 , for an index T' of 3,000 words the classification rate is around 75% which corresponds to a loss of around 5% on the classification rate with an important reduction rate of about 96%. By decreasing order of performance, $ECCD$ criterion obtains the best results, followed by GSS , IG , CC , DFL , DGL and χ^2 . Globally, these results confirm the different comparative studies performed on smaller collections. We can notice that the χ^2 criterion leads to unsatisfactory results when, according to [21], those results should have been similar to those obtained with IG . The $ECCD$ criterion leads to the best results and this improvement is more significant for a small index size. For an index size of 5,495 words, which corresponds to a reduction rate of 92.93%, the classification rate is 77.86%, which corresponds to a loss of only 1.18%.

6. CONCLUSION

This study, realized on the INEX XML Mining 2008 collection, confirms the importance in the use of selection feature for the categorization of XML documents and emphasizes the efficiency of criteria derived from the χ^2 such as GSS . It also shows the interest of using the words frequency with the entropy. Indeed, we observed that taking into account the occurrence of terms within the different classes using the entropy of the terms lets us improve significantly classification rate. We can also note that the term occurrence could be integrated into the other criteria. To this extend, we could calculate these criteria with the number of term per class instead of the number of document that contain the term.

7. REFERENCES

- [1] M. F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A. G. Chin, editor, *Text Databases and Document Management: Theory and Practice*, pages 78–102. Idea Group Publishing, Hershey, US, 2001.
- [2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

³Lemur Project: <http://www.lemurproject.org>

⁴Liblinear: <http://www.csie.ntu.edu.tw/~cjlin/liblinear/> - L2 loss support vector machine primal

- [3] L. Denoyer and P. Gallinari. The Wikipedia XML corpus. *SIGIR Forum*, 40(1):64–69, 2006.
- [4] L. Denoyer and P. Gallinari. Overview of the INEX 2008 XML Mining Track. In *Proceedings of the INEX Workshop Initiative for Evaluation of XML Retrieval*, pages 401–411, 2008.
- [5] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM'98: Proceedings of the 7th international conference on Information and knowledge management*, pages 148–155, New York, NY, USA, 1998. ACM.
- [6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [7] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.
- [8] L. Galavotti, F. Sebastiani, and M. Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 59–68. Springer-Verlag, 2000.
- [9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques, 2nd edition*. Morgan Kaufman Publishers, 2006.
- [10] B. C. How and W. T. Kiong. An examination of feature selection frameworks in text categorization. In *AIRS'05: Proceedings of 2nd Asia information retrieval symposium*, pages 558–564. Lecture notes in computer science, 2005.
- [11] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveiol, editors, *ECML'98: Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag, Heidelberg, DE, 1998.
- [12] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the Speech and Natural Language Workshop*, pages 212–217. Defense Advanced Research Projects Agency, Morgan Kaufmann, 1992.
- [13] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *SDAIR'94: Proceedings of the Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994.
- [14] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- [15] Y. H. Li and A. K. Jain. Classification of text documents. *The Computer Journal*, 41:537–546, 1998.
- [16] I. Moulinier and J.-G. Ganascia. Applying an existing machine learning algorithm to text categorization. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 343–354. Springer-Verlag, 1996.
- [17] H. T. Ng, W. B. Goh, and K. L. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–73, 1997.
- [18] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [19] J. S. Ronen Feldman. *The text mining handbook : Advanced approaches to analysing unstructured data*. Cambridge University Press, Cambridge, 2007.
- [20] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [21] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.
- [22] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.
- [23] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [24] E. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In *SDAIR'95: Proceedings of the 4th Symposium on Document Analysis and Information Retrieval*, pages 317–332, 1995.
- [25] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR'99: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, 1999.
- [26] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *ICML'97: Proceedings of the 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers, San Francisco, US, 1997.